SAMMI: Segment Anything Model for Malaria Identification

Luca Zedda, Andrea Loddo and Cecilia Di Ruberto
Department of Mathematics and Computer Science, University of Cagliari, Italy

Keywords: Computer Vision, Deep Learning, Object Detection, Image Processing, Blood Smear Images, Malaria Parasite Detection.

Abstract: Malaria, a life-threatening disease caused by the Plasmodium parasite, is a pressing global health challenge. Timely detection is critical for effective treatment. This paper introduces a novel computer-aided diagnosis system for detecting Plasmodium parasites in blood smear images, aiming to enhance automation and accessibility in comprehensive screening scenarios. Our approach integrates the Segment Anything Model for precise unsupervised parasite detection. It then employs a deep learning framework, combining Convolutional Neural Networks and Vision Transformer to accurately classify malaria-infected cells. We rigorously evaluate our system using the IML public dataset and compare its performance against various off-the-shelf object detectors. The results underscore the efficacy of our method, demonstrating superior accuracy in detecting and classifying malaria-infected cells. This innovative Computer-aided diagnosis system presents a reliable and near real-time solution for malaria diagnosis, offering significant potential for widespread implementation in healthcare settings. By automating the diagnosis process and ensuring high accuracy, our system can contribute to timely interventions, thereby advancing the fight against malaria globally.

1 INTRODUCTION

Malaria is a deadly disease caused by the Plasmodium parasite, and it continues to pose a significant public health challenge worldwide, with a high number of cases and fatalities. According to recent statistics from the WHO, there were approximately 247 million malaria cases and 619,000 deaths in 2021 (WHO, 2022). Most of these cases and fatalities occurred in Africa, with young children being the most vulnerable group. The disease is primarily spread through the bites of infected female Anopheles mosquitoes, and it affects red blood cells (RBCs) in humans, causing a range of symptoms and complications.

There are five species of the Plasmodium parasite, which are Falciparum (*Pf*), Vivax (*Pv*), Ovale (*Po*), Malariae (*Pm*), and Knowlesi (*Pk*). Among these, Pf is currently the most lethal for humans and is responsible for causing most malaria-related deaths. On the other hand, P. Vivax and P. Ovale are less harmful, but they can remain dormant for months in the liver and then reactivate, leading to acute respiratory distress syndrome. Pm can remain inactive in the blood for several years, whereas Pk has a shorter cycle and is the least fatal of all the species.

Detecting malaria as early as possible is crucial for quick treatment and management. Various diagnostic techniques can be used to diagnose malaria, including microscopical analysis of blood smears, rapid diagnostic tests, or real-time polymerase chain reaction. However, microscopy remains the most preferred method for diagnosing malaria due to its sensitivity, affordability, and ability to identify parasite species and density. Nevertheless, microscopy has several drawbacks, such as the need for highly experienced microscopists, limited access to this diagnostic method in some rural health facilities, and misdiagnosis due to low parasitemia or mixed infections.

In this context, Computer-aided diagnosis (CAD) systems can provide a viable solution to these challenges by assisting pathologists in diagnosing diseases and monitoring therapy.

This paper proposes a reliable and novel CAD system for detecting Pv parasites in blood smear images. The proposed system utilizes FastSAM for image segmentation and a deep learning approach based on convolutional neural networks (CNNs) and vision transformers (ViTs) for cell classification. The system aims to automate malaria diagnosis and improve accessibility in comprehensive screening scenarios. Identifying tiny parasites in near real-time enables the detection of different malaria species and a successive classification of the various life stages.

The performance of the proposed CAD system

Zedda, L., Loddo, A. and Di Ruberto, C.
SAMMI: Segment Anything Model for Malaria Identification.
DOI: 10.5220/0012325500003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 367-374
ISBN: 978-989-758-679-8; ISSN: 2184-4321

has been evaluated using the publicly available IML dataset and compared with existing object detectors. The results demonstrate the effectiveness of our approach in accurately detecting and classifying malaria-infected cells. Our proposed system contributes to improved diagnosis and management of malaria by leveraging advancements in deep learning techniques.

The main contributions of our work are listed as follows:

1. We have designed a novel pipeline for detecting Pv parasites in blood smear images.
2. We propose a novel technique to exploit the segmentation capabilities over an unseen domain.
3. We propose a classification approach studied for a high parasite-to-RBC imbalance scenario.
4. We define a classification approach for the parasites' life stages classification.

The rest of this work is structured as follows. Section 2 describes the current state of the art for malaria detection and stage classification, Section 3 summarizes the used materials and methods, Section 4 describes the experimental setup, evaluation and results. The paper concludes with Section 5, which overviews the work and draws conclusions proposing possible future outcomes.

2 RELATED WORK

Malaria detection remains a challenging task, especially in resource-limited settings. Microscopic examination is the gold standard for malaria diagnosis. Still, it is subjective and prone to errors as it involves trained microscopists manually inspecting blood smears to identify and classify malaria parasites based on their morphology. In recent years, computer-vision-based methods have achieved state-of-the-art results on malaria detection tasks.

In recent years, computer-vision-based methods have become popular for automated malaria detection. For example, CNNs are powerful image analysis tools that can classify malaria-infected cells by learning features from raw images (Zedda et al., 2022).

Object detection models, such as You Only Look Once (YOLO) and Faster R-CNN (FRCNN), have been employed for precise localization and identification of malaria parasites within blood smear images. These models can detect multiple parasites simultaneously and provide bounding box annotations, aiding in accurate diagnosis (Sultani et al., 2022).

Morphology and texture analysis techniques, rooted in traditional image processing, have also been utilized for malaria detection. These methods extract relevant features from images and employ machine learning algorithms for classification. By capturing distinctive morphological and textural characteristics of malaria parasites, these techniques contribute to accurate detection (Loddo et al., 2018).

Transfer learning has proven to be an effective strategy for malaria detection. Transfer learning enables high accuracy and efficiency in malaria detection tasks by leveraging pre-trained models on large-scale datasets, such as ImageNet, and fine-tuning them on malaria-specific datasets. This approach benefits from the learned representations of general image features and adapts them to the specific context of malaria detection (Loh et al., 2021).

Several recent studies have utilized the IML dataset for segmenting and classifying malaria-infected cells. For instance, Arshad et al. (Arshad et al., 2022) employed the ResNet50v2, achieving an accuracy of 95.63%, while Sengan et al. (Sengar et al., 2022) utilized Vision Transformer, achieving 90.03% for malaria detection. In addition, Mukherjee et al. (Mukherjee et al., 2021) achieved a Dice score of 95.0% for segmenting malaria-infected cells with a CNN-based approach. These works demonstrated the effectiveness of deep learning-based models in malaria parasite analysis on the IML dataset.

3 MATERIALS AND METHODS

3.1 Dataset

The IML dataset (Arshad et al., 2022) consists of 345 images of blood samples taken from individuals infected with P. Vivax malaria in Pakistan's Punjab province. Each image contains approximately 111 blood cells and includes accurate labels indicating the life stages and red blood cells. Figure 1 provides various examples of these full-size samples. The dataset encompasses four parasite life stages: *ring* (164 samples), *trophozoite* (77 samples), *schizont* (27 samples), and *gametocyte* (261 samples). A visual representation of these stages can be seen in Figure 2.

The images have a resolution of 1280×960 pixels and a 24-bit color depth, captured using a microscope-attached camera magnified at 100x.

3.2 Convolutional Neural Networks

CNNs excel in image classification and object detection by learning spatial features and handling large datasets. They comprise multiple layers, including convolutional, pooling, and fully connected layers.

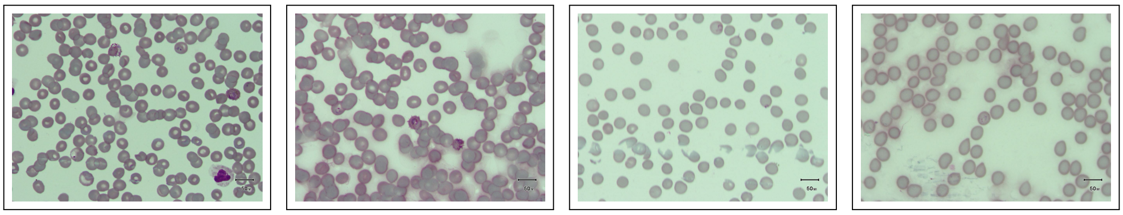
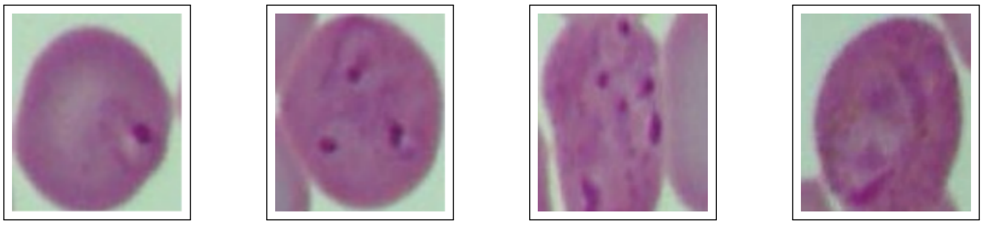


Figure 1: Samples of the full-size images contained in IML.



(a) Ring. (b) Trophozoite. (c) Schizont. (d) Gametocyte.

Figure 2: Life stages of malaria parasites contained in IML. From left to right: ring, trophozoite, schizont, and gametocyte stage.

CNNs detect simple features like edges in lower layers and complex features like object parts in higher layers. Popular architectures include ResNet, Inception, and ConvNext. CNNs enable applications like self-driving cars and medical image analysis. Ongoing research explores innovations like attention mechanisms (Woo et al., 2018) and generative models (Croitoru et al., 2023)..

3.3 Object Detectors and YOLO

Object detection methods rely on CNNs and are categorized into one-stage and two-stage architectures. Two-stage architectures, such as FRCNN, extract regions of interest, followed by classification and bounding box regression. One-stage detectors, such as RetinaNet and YOLO, directly generate bounding boxes and classes from predetermined anchors. These detectors are faster and better suited for time-sensitive applications and devices with computational constraints (Zou et al., 2023).

The YOLO family employs an end-to-end differentiable network that integrates bounding box estimation and object identification. YOLO divides the input image into a fixed grid and predicts bounding boxes and classes for each grid. YOLO is renowned for its speed and has been utilized for real-time object detection (Wang et al., 2022). in self-driving cars and surveillance systems (Narejo et al., 2021).

3.4 Vision Transformer

ViT uses a transformer encoder instead of standard convolutions (Khan et al., 2022; Dosovitskiy et al., 2021). It performs image classification in two phases: feature extraction and classification. In the feature extraction phase, the original image is transformed into a 1D sequence of patches, which undergo a linear projection and combine 1D position embedding with patch embeddings. The attention mechanism is a significant advancement in computer vision tasks, particularly with the evolution of transformer architectures and multi-head self-attention (MHSA) (Vaswani et al., 2017). ViT is more effective than traditional convolutional neural networks in capturing long-range dependencies and modeling global image features. However, processing large images with many patches can be computationally expensive. Several techniques have been proposed to address this issue, such as using overlapping patches or hierarchical patch representations for the SwinTransformer (Liu et al., 2021; Liu et al., 2022a).

3.5 The Proposed Pipeline

The Segment Anything (SA) project introduces a novel approach to image segmentation, including a promptable model called *SAM* and a large dataset, *SA-1B*. SAM is designed to transfer knowledge to new image distributions and tasks without additional training. It achieves remarkable zero-shot performance, rivaling or surpassing fully supervised models. The promptable segmentation task focuses on generating accurate segmentation masks given any prompt, even in cases of ambiguity. SAM's architecture, composed of an image encoder, prompt encoder, and mask decoder, enables real-time mask generation and improves efficiency through prompt reuse and cost amortization (Kirillov et al., 2023).

FastSAM (Zhao et al., 2023) proposes a CNN-based detector and prompt-guided selection in two stages to tackle the real-time constraint. It achieves comparable performance to SAM while reducing computational demands by utilizing the SA-1B dataset and YOLOv8 appropriately adapted for the segmentation task, namely YOLOv8-seg, FastSAM.

This paper proposes a novel pipeline for object segmentation and subsequent classification for parasite detection on blood smear images. The pipeline employs FastSAM as the region proposal extractor and ConvNext-small (Liu et al., 2022b) model for object classification. The pipeline diagram is depicted in Figure 3.

3.6 Metrics

Classification Metrics. To evaluate the models' performance in classification, we considered several metrics: *accuracy*, *precision*, *sensitivity*, and *F1-score*, which are defined below.

The classification outcome influences the following values for a given instance:

- *True Negatives (TN)*: Instances belonging to the *negative* class that were correctly predicted.
- *False Positives (FP)*: Instances belonging to the *negative* class that were incorrectly predicted.
- *False Negatives (FN)*: Instances belonging to the *positive* class that were incorrectly predicted.
- *True Positives (TP)*: Instances belonging to the *positive* class that were correctly predicted.

Accuracy (Acc) (see Equation (1)) measures the overall correctness of the model's predictions by calculating the ratio of correctly classified samples to the total number of samples. It is expressed as:

$$Accuracy = \frac{TP + TF}{TP + TF + FP + FN} \tag{1}$$

Sensitivity (Sen), or Recall (Rec) (see Equation (2)) measures the ability of the classifier to predict the positive class against FN:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

Precision (Pre) (see Equation (3)) measures the positive instances correctly classified among all instances classified as positive:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

F-Measure (F1) (see Equation (4)) (F1) provides a balanced evaluation by considering both false positives and false negatives. The F-Measure is calculated using the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{Pre \cdot Rec}{Pre + Rec} \tag{4}$$

Also, we used the macro average as we deal with an unbalanced dataset, and the number of samples in different classes varies significantly. It calculates the metric for each class separately and then takes the average, providing a balanced assessment of the model's performance across all classes.

Detection Metrics. Object detection methods are commonly assessed using the mean *average precision* (AP) metric and its variations (Lin et al., 2015). Precision relies on the Intersection over Union (IoU) concept to gauge detection accuracy. Specifically, IoU measures the ratio of the overlap area between the predicted bounding box and the actual object relative to the combined total area.

If the IoU surpasses a specific threshold, the detection is accurate and categorized as a TP. Conversely, the detection is labeled an FP if the IoU falls below the threshold. Moreover, if the model fails to detect an object in the ground truth, it is termed a FN.

As *precision* for object detection is defined in the same way as the classification one (see Equation (3)), the experimental evaluations were conducted considering five variants of the mAP metric:

- **AP** is evaluated with 10 different IOUs varying in a range of 50% to 95% with steps of 5%;
- $\mathbf{AP}_{50}$ is evaluated with a single values of IOU corresponding to 50%;
- $\mathbf{AP}_{\mathrm{s}}$ is the AP determined for small objects (with area $< 32^2$ pixels);
- $\mathbf{AP}_{\mathrm{m}}$ is the AP determined for medium objects (with $32^2 <$ area $< 96^2$ pixels);
- $\mathbf{AP}_{\mathrm{L}}$ is the AP determined for large objects (with area $> 96^2$ pixels).

average recall (AR) is another widely used metric in object detection, calculating recall values across various IoU thresholds, akin to AP. For consistency, we assess AR using identical IoU steps as AP, ranging from 50% to 95% with 5% increments, ensuring clear and coherent evaluation.

4 EXPERIMENTAL EVALUATION

4.1 Experimental Setup

The experiments were conducted on a desktop PC equipped with the following hardware specifications:

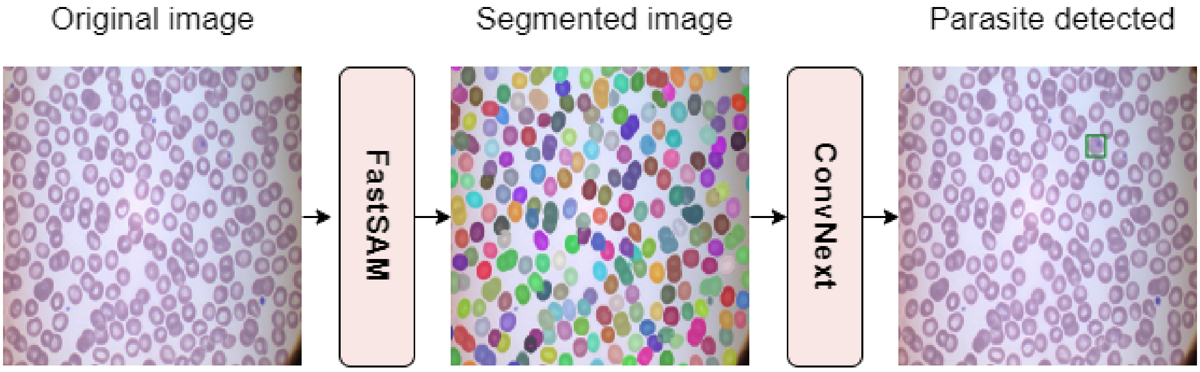


Figure 3: Pipeline visualization using ConvNext as the parasite discriminator.

Table 1: Image augmentation setup.

Augmentation	Parameters	Probability (%)
Horizontal Flip	-	50
Vertical Flip	-	50
Random Rotate 90	-	50
Random Resized Crop	Size: (*img_size*, *img_size*), Scale: (0.5, 1.0)	50
Grid Distortion	Num Steps: 5, Distort Limit: 0.3	50

an *Intel(R) Core(TM) i5-9400f* CPU operating at *4.1GHz*, *32GB* RAM, and an *NVIDIA RTX 3060 GPU* with *12GB* memory.

Training Details. We opted for a confidence threshold of 0.1 for FastSAM predictions and an IoU threshold of 0.8 for NMS. Two distinct classification architectures, namely ViT-base and ConvNext-base, were utilized. All networks were initialized using pre-trained weights from the ImageNet dataset (Deng et al., 2009). For optimization, we employed the *AdamW* optimizer with a learning rate of $1e-4$ and weight decay of 0.001. Each classification model underwent training for 100 epochs, utilizing a batch size of 32.

The original IML splits were used for parasite extraction using FastSAM and life stage classification. Additionally, we extracted 10% of the original training set to create a validation set. The best-performing model, determined based on cross-entropy loss using the validation set, was chosen as the reference for evaluation. We utilized various YOLO versions and sizes to assess the pipeline's performance, specifically YOLOv5 and YOLOv8 with medium and large-sized models. The detection models for comparison were trained for 50 epochs with the default YOLOv8 Ultralytics parameters [1]https://github.com/ultralytics/ultralytics.

To ensure the stability and significance of our method, the experiments were repeated five times, with the starting seed changed at each iteration.

Data Selection and Preparation. Considering the large number of predicted FastSAM structures, including red and white blood cells and cell clumps, we mitigated the imbalance issue by selecting 25 random non-parasitized structures as negative examples. Additionally, we implemented several augmentation techniques to balance the number of parasites with the negative examples. Details of the augmentations employed are provided in Table 1.

4.2 Experimental Results

Results on Malaria Parasites Detection. The outcomes of the detection experiments are outlined in Table 2. Despite possessing lower AP values, our approach yielded superior AP$_{50}$ results. Notably, YOLO-based object detectors were trained explicitly for malaria detection, while FastSAM operates without requiring any training, facilitating an unsupervised structure detection phase within the adopted domain. Additionally, our classifier training phase works on reduced image sizes compared to the original full-size images, enabling fast fitting and a modular discriminator.

However, the proposed methodology pipeline exhibited a recall value of only 0.51, which is inferior to fully supervised methods. This discrepancy can be attributed primarily to the presence of clumps formed by healthy and parasitized red blood cells. Future investigations should devise effective preprocessing

[1] Glenn Jocher, Ayush Chaurasia, and Jing Qiu, YOLO by Ultralytics, version 8.0.0 (accessed on October 9, 2023

Table 2: Experimental detection results obtained on the IML dataset (Arshad et al., 2022). The reported performance metrics include AP and AR at different IoU thresholds and AP at different scales. The number of parameters for each model is also provided. The best results are indicated in bold.

Model	**AP**	**AP**$_{50}$	**AP**$_{s}$	**AP**$_{m}$	**AP**$_{L}$	**AR**	**Params (M)**
YOLOv5m	0.52	0.62	-	0.38	0.56	0.66	21
YOLOv8m	**0.54**	0.60	-	**0.46**	**0.57**	**0.67**	**26**
YOLOv5l	0.49	0.56	-	0.35	0.54	0.65	47
YOLOv8l	0.52	0.60	-	0.43	0.56	0.66	44
Our method (w. ConvNext)	0.40	**0.85**	-	0.31	0.43	0.51	27
Our method (w. ViT)	0.39	0.81	-	0.29	0.41	0.52	86

Table 3: Experimental results are presented for the stage classification task on the IML dataset (Arshad et al., 2022). Original crops are utilized as training samples, and the evaluation is performed on detected test set crops. The best results for every type of architecture are indicated in bold.

Model	**Accuracy**	**F1-score**	**Precision**	**Sensitivity**	**Params (M)**
internimage-t	**0.65**	**0.56**	**0.56**	**0.62**	**29**
internimage-s	0.30	0.21	0.25	0.21	50
dino-vits16	0.54	0.42	0.44	0.41	**21**
dino-vitb16	**0.76**	**0.57**	**0.56**	**0.58**	85
convnextv2-tiny	0.71	0.54	**0.56**	0.54	**27**
convnextv2-base	**0.75**	**0.56**	0.56	**0.57**	87
swinv2-tiny	0.74	0.56	0.55	0.56	**27**
swinv2-base	**0.75**	**0.56**	**0.59**	**0.58**	86
vit-base	0.76	0.73	0.81	0.69	**86**
vit-large	**0.80**	**0.76**	**0.85**	**0.73**	303

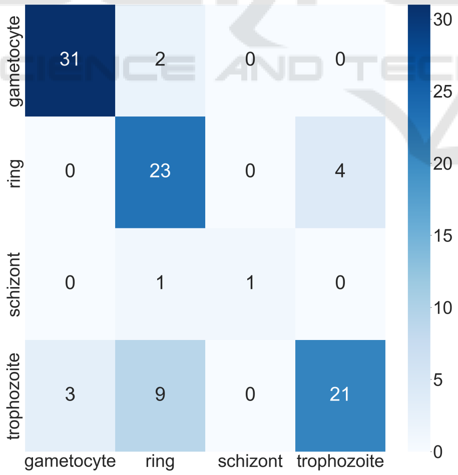


Figure 4: Confusion matrix illustrating the results of malaria parasite life stage classification. Rows represent the actual life stages (ring, trophozoite, schizont, gametocyte), while columns indicate the predicted stages.

strategies tailored to address this challenge.

In terms of training time, YOLO-based architectures demand 10 GPU/hours for medium-sized models and 20 GPU/hours for large-sized ones. In contrast, our discriminator necessitates only 1 GPU/hour of training, making it significantly faster and more lightweight. These timescales accelerate the training pace, facilitate rapid experimentation, and enable further studies, even on computationally limited machines.

Results on the Life Stage Classification. For fairness, multiple classification models were trained using the detected parasites, allowing for classification across various parasite stages. The best-performing model was evaluated on the validation set, employing methodologies consistent with those used for the parasite discriminator. More precisely, based on its superior AP, the classification analysis was conducted on the crops extracted from the top-performing detection model, specifically YOLOv8m. The comprehensive experimental results are summarized in Table 3.

The outcomes reveal that the vit-large model achieved the highest F-measure, reaching 76.45%, despite the considerable imbalance observed in the distribution of different stages, as expressed in Section 3.1. The confusion matrix of the vit-large model presented in Figure 4 underscores the challenges associated with accurate classification, particularly concerning the ring and schizont classes. These classes exhibit morphological similarities, posing significant difficulties even for domain experts.

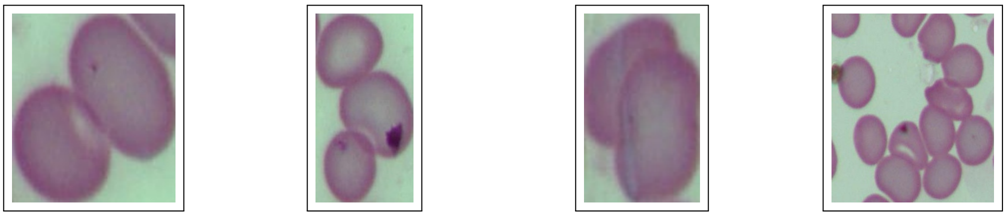


Figure 5: Examples of clumps extracted by FastSAM.

4.3 Discussion

Real-Time Analysis. The real-time capabilities of the proposed pipelines were evaluated using our designated test set. For each image within, the Frames Per Second (FPS) was computed by averaging the time taken in milliseconds for the complete parasitized cells proposal and filtering process. On average, the pipeline integrating FastSAM and ConvNext processed each image in 82 milliseconds (equivalent to 12 FPS), while the ViT-based pipeline required 86 milliseconds per image (equivalent to 11 FPS) due to its larger parameter count. The processing time in milliseconds for each image displayed minimal variance, typically falling within the range of 350 to 400 proposed regions.

Limitations. While our innovative pipeline showcased the potential of the FastSAM architecture in malaria detection, it warrants further investigation. The experiments revealed challenges arising from closely juxtaposed red blood cells forming clumps, leading the models to inaccurately predict them as singular objects. These challenges, as depicted in Figure 5, can be attributed to two primary factors: the utilization of FastSAM without a fine-tuning strategy and the inherent issue of clumps, which requires specific post-processing techniques.

5 CONCLUSIONS

Despite the issues in identifying cell clumps, the proposed malaria detection pipeline, utilizing FastSAM for object proposal extraction and a subsequent life stage classification phase composed of ConvNext, demonstrates that the FastSAM architecture is applicable in malaria detection in a semi-supervised context.

It exhibited remarkable versatility. Beyond its primary purpose in malaria diagnosis, this pipeline can be tailored for diverse medical imaging tasks, from identifying and classifying different types of blood cells, including white blood cells (leucocytes) and fragmented red blood cells (schistocytes). This scenario can showcase its adaptability in diagnosing blood disorders, infections, and conditions.

The first step of future research is to enrich the analysis of the segmented clumps by estimating the number of RBCs. Then, our aim is to study the impact of different preprocessing steps on full-size images to improve detection results. Also, the stage classification results showed impressive results despite the unbalanced stages. These results may also be increased by applying preprocessing to the crops to emphasize the parasitic structures on the inside of the red blood cells.

Aside from exploring preprocessing techniques on the data, we will also test our novel pipeline on different datasets to validate our approach. Finally, we plan to adopt a non-supervised approach for parasite discrimination, allowing for fully non-supervised malaria detection.

Funding

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5—Call for tender No. 3277 published on December 30, 2021, by the Italian Ministry of University and Research (MUR) funded by the European Union—NextGenerationEU. Project Code ECS0000038—Project Title eINS Ecosystem of Innovation for Next Generation Sardinia—CUP F53C22000430001—Grant Assignment Decree No. 1056 adopted on June 23, 2022, by the Italian Ministry of University and Research (MUR).

REFERENCES

Arshad, Q. A., Ali, M., Hassan, S., Chen, C., Imran, A., Rasul, G., and Sultani, W. (2022). A dataset and benchmark for malaria life-cycle classification in thin blood smear images. *Neural Comput. Appl.*, 34(6):4473–4485.

Croitoru, F., Hondru, V., Ionescu, R. T., and Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):10850–10869.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Khan, S. H., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022). Transformers in vision: A survey. *ACM Comput. Surv.*, 54(10s):200:1–200:41.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W., Dollár, P., and Girshick, R. B. (2023). Segment anything. *CoRR*, abs/2304.02643.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft coco: Common objects in context.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B. (2022a). Swin transformer V2: scaling up capacity and resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11999–12009. IEEE.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., and Xie, S. (2022b). A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11966–11976. IEEE.

Loddo, A., Ruberto, C. D., and Kocher, M. (2018). Recent advances of malaria parasites detection systems based on mathematical morphology. *Sensors*, 18(2):513.

Loh, D. R., Yong, W. X., Yapeter, J., Subburaj, K., and Chandramohanadas, R. (2021). A deep learning approach to the screening of malaria infection: Automated and rapid cell counting, object detection and instance segmentation using mask R-CNN. *Comput. Medical Imaging Graph.*, 88:101845.

Mukherjee, S., Chatterjee, S., Bandyopadhyay, O., and Biswas, A. (2021). Detection of malaria parasites in thin blood smears using cnn-based approach. In Mandal, J. K., Mukherjee, I., Bakshi, S., Chatterji, S., and Sa, P. K., editors, *Computational Intelligence and Machine Learning*, pages 19–27, Singapore. Springer Singapore.

Narejo, S., Pandey, B., Esenarro Vargas, D., Rodriguez, C., and Anjum, M. (2021). Weapon detection using yolo v3 for smart surveillance system. *Mathematical Problems in Engineering*, 2021:1–9.

Sengar, N., Burget, R., and Dutta, M. (2022). A vision transformer based approach for analysis of plasmodium vivax life cycle for malaria prediction using thin blood smear microscopic images. *Computer Methods and Programs in Biomedicine*, 224:106996.

Sultani, W., Nawaz, W., Javed, S., Danish, M. S., Saadia, A., and Ali, M. (2022). Towards low-cost and efficient malaria detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20655–20664. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Wang, C., Bochkovskiy, A., and Liao, H. M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *CoRR*, abs/2207.02696.

WHO, W. H. O. (2022). World Malaria Report 2022.

Woo, S., Park, J., Lee, J., and Kweon, I. S. (2018). CBAM: convolutional block attention module. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 3–19. Springer.

Zedda, L., Loddo, A., and Di Ruberto, C. (2022). A deep learning based framework for malaria diagnosis on high variation data set. In *Image Analysis and Processing - ICIAP 2022 - 21st International Conference, Lecce, Italy, May 23-27, 2022, Proceedings, Part II*, volume 13232 of *Lecture Notes in Computer Science*, pages 358–370. Springer.

Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., and Wang, J. (2023). Fast Segment Anything.

Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J. (2023). Object detection in 20 years: A survey. *Proc. IEEE*, 111(3):257–276.